JSLHR

Tutorial

Masked Visual Analysis: Minimizing Type I Error in Visually Guided Single-Case Design for Communication Disorders

Tara McAllister Byun,[a] Elaine R. Hitchcock,[b] and John Ferron[c]

Purpose: Single-case experimental designs are widely used to study interventions for communication disorders. Traditionally, single-case experiments follow a response-guided approach, where design decisions during the study are based on participants' observed patterns of behavior. However, this approach has been criticized for its high rate of Type I error. In masked visual analysis (MVA), response-guided decisions are made by a researcher who is blinded to participants' identities and treatment assignments. MVA also makes it possible to conduct a hypothesis test assessing the significance of treatment effects.
Method: This tutorial describes the principles of MVA, including both how experiments can be set up and how results can be used for hypothesis testing. We then report a case study showing how MVA was deployed in a multiple-baseline across-subjects study investigating treatment for residual errors affecting rhotics. Strengths and weaknesses of MVA are discussed.
Conclusions: Given their important role in the evidence base that informs clinical decision making, it is critical for single-case experimental studies to be conducted in a way that allows researchers to draw valid inferences. As a method that can increase the rigor of single-case studies while preserving the benefits of a response-guided approach, MVA warrants expanded attention from researchers in communication disorders.

Single-case design, also called single-subject experimental design, is a widely used methodology in the study of interventions for communication disorders (Kearns, 1986; McReynolds & Kearns, 1983; McReynolds & Thompson, 1986; Robey, Schultz, Crawford, & Sinner, 1999; Thompson, 2006). The single-case design involves measuring a behavior before, during, and/or after treatment in an individual or small sample of individuals. A distinguishing characteristic of single-case design is the collection of multiple repeated measurements of a dependent variable within each subject (Horner & Odom, 2014; Thompson, 2006). Using these repeated measures, the design aims to demonstrate a functional or causal relationship between the application of an intervention and a change in the dependent variable. Replication of this demonstration either within or across participants is essential to control for threats to internal validity and/or establish the robustness of the effect. The single-case design should not be confused with the similarly named case study design, which lacks the element of experimental control.

Although the randomized controlled trial (RCT) continues to be regarded as the gold standard for establishing the efficacy of a treatment, the RCT methodology is not always feasible, particularly if the condition under investigation has a low prevalence in the population (Byiers, Reichle, & Symons, 2012; Kazdin, 2010). In addition, by collecting detailed data on individual participants over time, single-case studies offer a more complete picture of the range of trajectories of response to an intervention than an RCT can (Howard, Best, & Nickels, 2015). Finally, in a multiphase model of clinical research (e.g., Robey, 2004), RCTs represent Phase III; this phase should be undertaken only after an initial phase of exploratory studies (Phase I) and a subsequent phase of small-scale systematic studies, including single-case designs (Phase II). By conducting careful single-case studies during Phase II, researchers can refine their intervention protocol, anticipate any confounding factors, and estimate an effect size so that they can conduct a power analysis to set the appropriate sample size for the RCT.

A number of different designs are included in the family of single-case studies. In the interest of brevity, the

[a]NYU Steinhardt School of Culture, Education, & Human Development
[b]Montclair State University, Bloomfield, New Jersey
[c]University of South Florida, Tampa
Correspondence to Tara McAllister Byun: tara.byun@nyu.edu

Editor: Julie Liss
Associate Editor: Tanya Eadie
Received August 30, 2016
Revision received December 5, 2016
Accepted January 20, 2017
https://doi.org/10.1044/2017_JSLHR-S-16-0344

Disclosure: *The authors have declared that no competing interests existed at the time of publication.*

present tutorial will focus on the multiple-baseline design, which is commonly used in research in communication disorders. Interested readers should see Byiers et al. (2012) or Horner and Odom (2014) for more comprehensive overviews. In a multiple-baseline study, treatment is applied in a staggered fashion to multiple participants or to multiple targets within one participant, such as different speech sounds in a child with a speech sound disorder. A schematic example of a multiple-baseline across-subjects design is presented in Figure 1. If the dependent variable remains stable throughout the baseline period and changes only after the application of treatment, and this effect is replicated across participants or targets, it provides evidence of a functional influence of treatment on the behavioral outcome.

Improving the Credibility of Single-Case Design

Although single-case design is well established in some fields, it is relatively unknown in others, which can pose a challenge when researchers communicate single-case findings to the broader research community (Kratochwill & Levin, 2014). In an effort to increase the rigor and reputation of single-case design, the What Works Clearinghouse (WWC) convened a panel of researchers to lay out consensus guidelines for acceptable design in the single-case methodology (Kratochwill et al., 2013). With respect to design, the WWC criteria require that the independent variable or treatment be systematically manipulated; the study cannot be observational or correlational. It is also specified that the study must include a minimum of three "phase transitions," that is, opportunities to demonstrate an intervention effect, at three different points in time. Additional criteria require a minimum number of observations in each phase (preferably no fewer than five, with a duration of three data points accepted "with reservations"), and a minimum level of interrater agreement (e.g., 80% point-to-point agreement for a categorical dependent variable) assessed over a minimum of 20% of data points in each phase of treatment.

Other efforts to strengthen the scientific reputation of single-case design have focused on how effects of treatment are assessed, and how decisions are made to undertake a transition from one phase of the study to another, such as from baseline to treatment. The traditional norm in single-case research uses a "response-guided" design, in which the experimenter inspects the data throughout the study and decides when to make transitions on the basis of visual impressions of the stability of the data (e.g., Thompson, 2006). The major advantage of this approach is that it allows the experimenter to minimize instances where the interpretation of results is complicated by "demonstrations of noneffect" such as rising baselines or excessive variability in the baseline phase. Although in some cases a rising baseline trend is a genuine reflection of maturation or other outside influences on participant performance, in other cases it may be a chance occurrence. In a response-guided approach, the researcher can opt to extend the baseline phase and observe whether stability is regained. When the transition from one phase to another is slated to occur after a predetermined number of sessions, this is not possible, and the researcher may be left uncertain whether improvements after the onset of intervention represent a true response to treatment or a continuation of a baseline trend. (An illustration is provided in Figure 2; note that if the initiation of treatment had occurred after the second data point, it would be unclear

Figure 1. Schematic example of multiple-baseline across-subjects data. The dotted line represents the point of introduction of biofeedback treatment. Adapted from "Investigating the use of traditional and spectral biofeedback approaches to intervention for /r/ misarticulation," by T. McAllister Byun and E. Hitchcock, 2012, American Journal of Speech-Language Pathology, 21, pp. 207–221. Copyright © 2012 by American Speech-Language-Hearing Association. Adapted with permission.

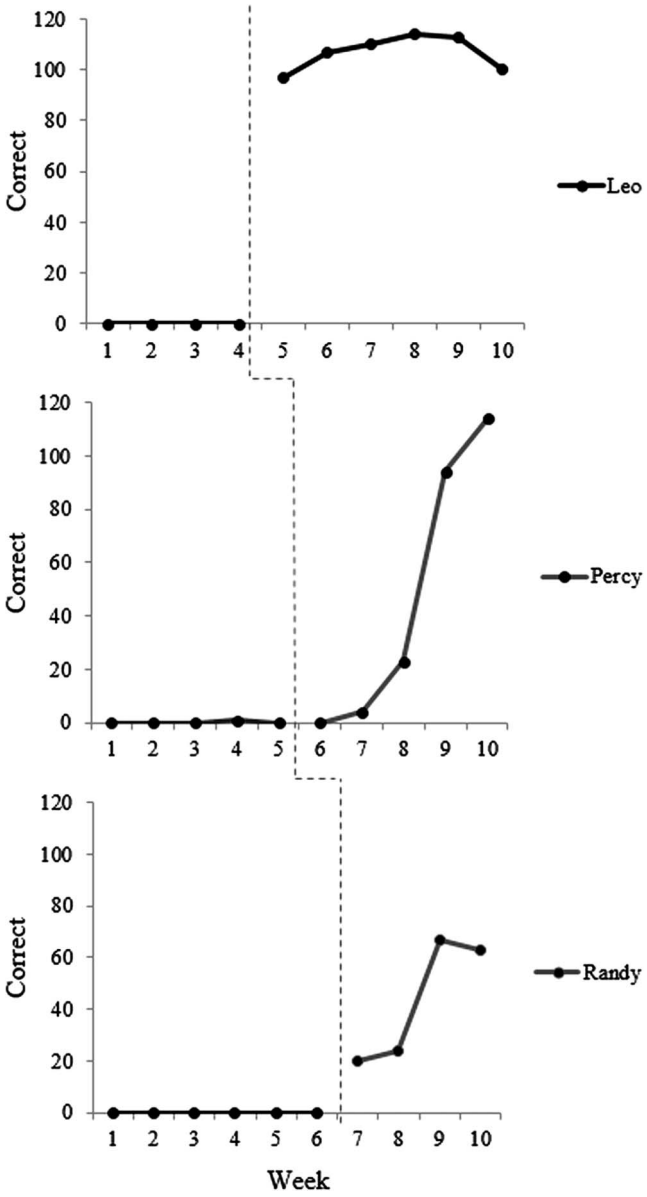
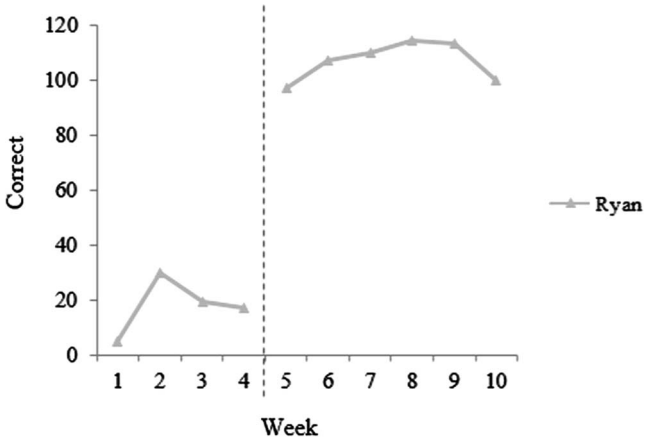


Figure 2. A rising or unstable baseline can complicate the interpretation of single-case data. Schematic example adapted from "Investigating the use of traditional and spectral biofeedback approaches to intervention for /r/ misarticulation," by T. McAllister Byun and E. Hitchcock, 2012, American Journal of Speech-Language Pathology, 21, pp. 207–221. Copyright © 2012 by American Speech-Language-Hearing Association. Adapted with permission.



whether subsequent gains reflected a response to treatment or a continuation of the baseline trend.) Because each participant in a single-case treatment study can represent a sizable investment of time on the part of the experimenter, many researchers are unwilling to risk compromising data interpretability by deviating from the response-guided design.

However, the response-guided design creates a different problem. When experimenters have the opportunity to choose when each participant will make the transition from baseline to treatment, they can take corrective action in response to high outliers or rising baselines; however, they may be more likely to accept or even favor low outliers and falling trends in the baseline. The principle of *regression to the mean* dictates that these unusually low points will typically be followed by a shift back to a rising trajectory, which may then be interpreted as a treatment effect. The response-guided approach is therefore associated with an increased likelihood of a Type I error, in which a spurious change is misinterpreted as evidence of a treatment effect (Ferron & Levin, 2014; Howard et al., 2015). Thus, the single-case methodology poses a double bind: Experimenters can use random assignment and risk difficulties with interpretation if data are compromised by demonstrations of noneffect, or they can adopt a response-guided approach and risk overestimating the magnitude of any treatment effect. The masked visual analysis (MVA) approach reviewed here (Ferron & Jones, 2006) is intended as a "best-of-both-worlds" solution that makes it possible to control the risk of Type I error while still undertaking phase transitions only when data are sufficiently stable. The MVA approach is described in detail in Section II; before proceeding, we briefly review methods used to measure treatment effects in single-case design.

Measuring Treatment Effects in Single-Case Design

Hand in hand with the response-guided approach to phase transitions, the traditional approach to single-case design draws on visual inspection of data points to document any effects of treatment. A treatment effect may be reported when the level, slope, and variability of data points remain stable within baseline phases but change each time there is a transition to treatment. However, visual inspection has been criticized as excessively prone to both Type I and Type II errors (Brossart, Parker, Olson, & Mahadevan, 2006; Howard et al., 2015; Matyas & Greenwood, 1990; Robey et al., 1999). Likewise, measures to quantify the amount of overlap in the values assumed by the dependent variable during treatment phases versus nontreatment phases, such as percentage of nonoverlapping data (Scruggs, Mastropieri, & Casto, 1987), have been criticized as yielding "unacceptably high levels of error" (Wolery, Busick, Reichow, & Barton, 2010: p. 18). Many single-case studies in communication disorders calculate effect sizes reflecting the magnitude of change over the course of intervention, often drawing on variants of Busk and Serlin's (1992) modification of Cohen's *d* (Beeson & Robey, 2006; Gierut & Morrisette, 2011; Gierut, Morrisette, & Dickinson, 2015). Effect size measures provide a valuable quantitative complement to visual inspection. However, they are limited by the lack of a standard reference for interpretation, because Cohen's (1988) benchmarks to interpret effect sizes in group designs are not valid when applied to single-case studies. When effect sizes are large, this may not pose a problem, but if effect sizes are small due to a slow rate of change and/or high variability, readers may disagree as to whether a given case represents a meaningful demonstration of effect. In such cases, it is highly desirable to have the ability to conduct a hypothesis test evaluating whether post-treatment performance represents a significant change relative to pretreatment levels.

Kratochwill and Levin (2010, 2014) have argued that the most effective strategy to bring single-case research into the scientific mainstream is to incorporate hypothesis tests that evaluate the statistical significance of any change observed, increasing the interpretability of results. To meet the assumptions for inferential statistics, they advocate the adoption of single-case designs incorporating randomization. The argument that randomization and hypothesis testing can elevate the status of single-case design in the broader scientific literature received an important endorsement in 2011 when the Oxford Center for Evidence-Based Medicine (OCEBM) added the *n-of-1 randomized controlled trial,* a type of single-case design with randomization, to its hierarchy of levels of evidence (OCEBM Levels of Evidence Working Group, 2011). Whereas the n-of-1 trial was absent from previous versions of the CEBM hierarchy, the 2011 hierarchy indicates that a series of n-of-1 trials can provide the highest level of evidence (Howard et al., 2015). In the communication disorders literature, the single-case randomization design has been promoted as an effective solution

in cases where a treatment effect coexists with a long-term learning or maturational trajectory (Rvachew, 1988). MVA makes it possible to incorporate randomization into a study design, legitimizing the use of statistical hypothesis testing, while avoiding some of the risks to interpretability that often arise in connection with randomization.

Masked Visual Analysis

MVA aims to reduce the influence of experimenter bias on single-case design while still using a response-guided approach to ensure that each phase of the study exhibits a consistent and interpretable response pattern. It also allows researchers to incorporate an element of randomization and thus make use of statistical hypothesis tests, but it does so without posing too great a risk that data will be compromised by demonstrations of noneffect. The MVA approach involves dividing the study team into two parts. The intervention team conducts the intervention and collects the data that will be used to measure treatment effects. The analysis team inspects the plotted data and selects an appropriate time to make a phase transition. However, the data reviewed by the analysis team are masked; that is, they do not carry information about the identity or treatment status of a participant. In the initial baseline phase, the analysis team looks for a point when all participants are maintaining a stable trajectory without problematic outliers or rising trends. In subsequent stages, the analysis team looks for evidence of a selective response to treatment. For example, in a multiple-baseline across-subjects study, the analysis team looks for one participant to show a response while other participants remain at baseline—although the analysis team remains unaware of the identity of any participant(s) in treatment. If a stable pattern has not yet emerged, the analysis team can request that additional data be collected before another phase transition is initiated. The analysis team's data-driven guesses about the masked aspects of the study can be used to conduct a hypothesis test and derive a *p* value, as described below. In the detailed discussion that follows, we continue with the example of a multiple-baseline across-subjects design, although MVA has in fact been extended to other single-case designs, including withdrawal and alternating treatment designs (for more information, see Ferron & Levin, 2014).

Setting the Initial Parameters

Before the initiation of any study activities, the intervention team and the analysis team must agree on the parameters that will guide the course of the study once underway. Because information that is accessible to one part of the team will be masked from the view of the other part of the team, it is essential to consider a comprehensive set of possibilities before initiating the study. To begin with, the team must agree on all of the elements that need to be operationalized for any treatment study. These include the number of participants to be enrolled, the number and duration of treatment sessions to be provided, the number of trials to be elicited per treatment session, the treatment targets and the contexts in which they will be elicited, the nature of cueing, and the nature of feedback to be provided (see, e.g., Maas et al., 2008). If the protocol calls for changes in treatment parameters in response to participant progress, such as reducing the level of clinician support or increasing the complexity of targets, these must also be agreed upon before the start of the study (see, e.g., Hitchcock & McAllister Byun, 2015; Rvachew & Brosseau-Lapré, 2012).

It is also necessary to set a number of parameters specific to the context of single-case experimental design incorporating MVA. A minimum number of observations per phase must be determined (e.g., a minimum of four baseline sessions must be collected from all participants before any participant can enter treatment). This may or may not be the same as the minimum stagger or spacing between participants or conditions after treatment has been initiated. Finally, the team should discuss what criteria will be used to identify a rising trend or classify an observation as an outlier, and they must also plan their response in the event that problematic data are observed (e.g., extend the phase by three sessions). Finally, experimenters may wish to set a contingency plan for cases in which a stable baseline is not maintained or a treatment effect is not observed: After a set number of sessions, the study can automatically proceed to the next phase, so all participants have a fair chance to receive treatment.

Role of the Intervention Team

The intervention team is responsible for delivering all treatment in accordance with the protocol specified during the parameter-setting phase. Performance during baseline, within-treatment, and/or maintenance sessions can be scored by the intervention team, who then share the scores with the analysis team in a masked format, that is, stripped of information that could identify participants or their treatment assignments. An alternative is to transmit raw session data, such as audio or video files, to a blinded third party who scores the data before transmitting them to the analysis team. This measure, discussed in more detail in the case study to follow, offers an additional level of protection against experimenter bias. Once the analysis team indicates that an adequately stable baseline has been established, the intervention team makes a random assignment pertaining to the provision of treatment. In a multiple-baseline design, this involves randomly selecting the participant or target that will begin to receive treatment. The intervention team informs the analysis team that a random assignment has been made, but the specific outcome is not conveyed. The intervention team then continues to provide intervention and transmit probe data to the analysis team until they receive the signal that it is time for an additional transition, at which point another random assignment will be made. This continues until the analysis team signals that the study should be ended, or until a predetermined study duration is reached.

Role of the Analysis Team

The analysis team plots and inspects the performance data that have been collected by the intervention team and scored by the intervention team or a third party. During the baseline phase, the analysis team looks for any problematic trends or outliers that could compromise experimental control. If such a deviation is noted, the analysis team requests that additional data points be collected before a phase transition is undertaken. Increasing the number of data points will give the analysis team a better basis for deciding whether an apparent rising trend is a robust effect or a random deviation, and it will diminish the influence of any outliers. Once all participants display a stable baseline, or once a predetermined time limit is reached, the analysis team indicates that a random determination about treatment initiation can be made. In a multiple-baseline across-subjects design, this involves transitioning one participant from a baseline phase to a treatment phase.

Once treatment has been initiated, the analysis team looks for a situation in which they can make a confident guess about the treatment status of all participants. In the first phase of a multiple-baseline across-subjects design, the analysis team looks for one participant to show evidence of a response to treatment while all others remain stable. (See the case study below for discussion of how this response to treatment might be operationalized.) If the analysis team is not able to establish a strong hypothesis as to which participant has advanced to treatment, they request that additional data points be collected prior to the next phase transition. This process is repeated as successive participants in the study make the transition from baseline to treatment. Finally, after a set number of phases or after a predetermined threshold for discontinuation is reached, the analysis team indicates that the study should be terminated.

Summative Analysis and Hypothesis Testing

The MVA procedure makes it possible to incorporate randomization into a single-subject design while reducing the risk that valuable data will be compromised by outliers or spurious trends. As discussed above, randomization has the benefit of increasing internal validity in single-case designs. It removes any opportunity for the experimenter's bias, on the basis of baseline performance, to influence how participants are assigned to treatment conditions. The second major benefit is that the incorporation of randomization and blinding makes it possible to conduct a hypothesis test and calculate a p value.

To conduct a hypothesis test in a study using MVA methodology, first the analysis team must make its best guess, on the basis of the complete set of plotted data, regarding the order of application of intervention in the study. In a multiple-baseline across-subjects design, the analysis team guesses the order in which participants made the transition from baseline to treatment, based on the sequence in which different participants' plotted values diverged from baseline levels. The null hypothesis holds that there is no effect of treatment, so treatment and no-treatment observations are not meaningfully different; if the null hypothesis is true, the analysis team is essentially making a random guess about the order of application of treatment. In this case, the probability of guessing the correct order purely by chance is equal to one divided by the total number of possible orderings in which participants could be transitioned from baseline to treatment. This value can be treated as an empirically derived p value, because the p value represents the probability of obtaining the observed outcome or a more extreme outcome purely by chance when the null hypothesis holds true. If the analysis team guesses the correct order of treatment transitions in a multiple-baseline across-subjects study with five participants, this probability is 1/5!, where 5! is the total number of possible orderings of five participants. We thus derive a p value of 1/120 = .008 and provisionally reject the null hypothesis of no treatment effect. If the analysis team guesses the correct intervention order not on the first but on the second try, the p value is 2/120 = .017; on the third try, p = 3/120 = .025; on the fourth try, p = 4/120 = .033; on the fifth try, p = 5/120 = .042; and on the sixth try, p = 6/120 = .05. Note that the intervention team does not disclose any additional information (such as how close the last guess was or which subjects have been ordered correctly) between guesses. Thus, even though the analysis team can make up to five guesses and still reject the null hypothesis with $p < .05$, it is not trivial to identify the correct order of treatment for all five participants within those first five attempts.

Case Study of Masked Visual Analysis in Speech-Language Intervention Rationale for Adopting Masked Visual Analysis

Two authors of this tutorial have collaborated for some time to investigate the efficacy of biofeedback interventions for residual or treatment-resistant misarticulation of the North American English /r/ sound (Hitchcock & McAllister Byun, 2015; Hitchcock, McAllister Byun, Swartz, & Lazarus, in press; McAllister Byun & Hitchcock, 2012; McAllister Byun, Hitchcock, & Swartz, 2014). This has given us ample opportunity to become familiar with several challenges faced by researchers investigating the treatment of residual speech errors. First, the prevalence of residual speech errors is low, estimated at less than 5% in the school-aged population (Shriberg, Tomblin, & McSweeny, 1999), and less than 2% in the college-aged population (Culton, 1986). Second, these residual errors are widely recognized to be challenging to treat (e.g., Ruscello, 1995); this is perhaps the main reason there is so much clinical and research interest in them. Given this combination of low prevalence and the need for a fairly extended period of treatment delivered by a skilled provider, it is difficult—though not impossible—to study the efficacy of interventions for residual speech errors using well-powered group RCTs. This has led us, along with many other researchers studying interventions for residual speech errors (e.g., Preston et al., 2014), to favor single-case experimental design in our treatment

studies. However, we have encountered challenges in that context as well. First, because our treatment is intended to produce lasting learning effects, we neither expect nor desire that gains made in treatment will disappear when the treatment is withdrawn; this rules out a reversal design. Second, the majority of children in our studies have no speech errors other than /r/ (although a handful do additionally present with dentalization or lateralization of /s/), which rules out a multiple-baseline across-behaviors design. If we treat /r/ in different positions or phonetic contexts (e.g., /ɝ/ versus /ɪɝ/) as different target behaviors, there is a high risk that progress in a treated context will generalize to the untreated context, compromising experimental control (e.g., McAllister Byun, Swartz, Halpin, Szeredi, & Maas, 2016). We are left with the multiple-baseline across-subjects design as our only major means of establishing experimental control. However, there is a challenge here as well: Participants typically do not respond immediately on application of treatment; a latency of three to five sessions (between 1 and 3 weeks) is more characteristic, but the duration of this lag is not predictable across individuals.[1] In addition, treatment for residual speech errors is not effective for 100% of clients, with most published studies including instances of nonresponders (e.g., McAllister Byun, Hitchcock, & Swartz, 2014; Preston, Brick, & Landi, 2013). With so many factors making it difficult to document a well-controlled effect of treatment, it is especially important to guard against any demonstrations of noneffect such as rising baselines or excessive variability in the baseline period. The MVA method provides this opportunity without increasing the risk that experimenter bias could influence our outcomes and without sacrificing the opportunity to incorporate randomization and conduct a hypothesis test.

Method

Treatment Parameters

The study, which is reported in greater detail in a companion article, consisted of three arms, where each arm was a multiple-baseline across-subjects study investigating the influence of one technology to deliver biofeedback intervention for residual errors affecting /r/.[2] One arm evaluated the effects of visual-acoustic biofeedback intervention (see McAllister Byun & Hitchcock, 2012; McAllister Byun et al., 2016); another tested ultrasound biofeedback (e.g., McAllister Byun, Hitchcock, & Swartz, 2014; Preston et al., 2013; Preston et al., 2014); and a final arm examined electropalatographic (EPG) biofeedback (e.g., Fletcher, Dagenais, & Critz-Crosby, 1991; Hitchcock et al., in press). Four participants with residual /r/ misarticulation were enrolled in each leg of the study. Participants ranged in age from 7;6 to 13;0 overall, with a mean age of 10;3 (*SD* = 21.7 months) in the visual-acoustic biofeedback arm, 9;9 (*SD* = 25.9 months) in the ultrasound biofeedback arm, and 8;3 (*SD* = 11.3 months) in the EPG biofeedback arm. One out of four participants in each arm was female, which is roughly consistent with previous research on the gender distribution of residual speech errors (Shriberg, 2010). During the intervention phase of the study, participants received individual biofeedback intervention from a certified speech-language pathologist in two 30-minute treatment sessions per week. The protocol for intervention, which was kept as consistent as possible across all three arms of the study, called for two initial instructional sessions in which participants were familiarized with the biofeedback display and coached on strategies for using the biofeedback to achieve a more accurate /r/ sound. After the first two sessions, all sessions began with a 5-minute pre-practice or "free play" period, followed by 72 trials in which /r/ was elicited at the syllable or word level. Trials were elicited in blocks of six; between blocks, the treating clinician would provide a qualitative comment serving as feedback for the previous block and/or providing the focus for the next block. In an effort to encourage generalization, the level of difficulty experienced within treatment was adjusted along an adaptive hierarchy on the basis of "challenge point" principles (Hitchcock & McAllister Byun, 2015; Rvachew & Brosseau-Lapré, 2012). The hierarchy for adjusting the difficulty of practice was predetermined and implemented via a customized software (*Challenge-R*; McAllister Byun, Hitchcock, & Ortiz, 2014).

Masked Visual Analysis Parameters

In each session, participants completed a standard probe measure eliciting 30 syllables, 25 words, and five sentences containing /r/ targets. In baseline and maintenance sessions, the probe was the only measure elicited. In treatment sessions, participants completed treatment activities as described above and then produced the same probe measures elicited in baseline and maintenance sessions. The decision to administer probes at the end of treatment sessions, when short-term gains could potentially affect probe performance, was intended to maximize the sensitivity of our measures to any response to treatment, in the interest of moving participants through phases of the study in a timely manner. For the same reason, the analysis team focused on vocalic /r/, which clinical and research evidence suggest to be the earliest-emerging variant (Klein, McAllister Byun, Davidson, & Grigos, 2013), in the simplest context (syllable level).

In this study, the intervention team and the analysis team were in physically different locations at Montclair State University and New York University. As noted above, speech tokens were collected and processed by the intervention team, but the tokens were rated by a third party prior to being plotted and inspected by the analysis team. Obtaining ratings from a blinded third party instead of the intervention

[1]The latencies described here refer to the typical length of time before gains become evident in generalization probes, which is commonly adopted as a clinically significant indication of progress. If performance were to be measured within the treatment setting, when clinician cues and/or biofeedback are available, more immediate gains may be evident.
[2]The data reported here have been adapted for clarity of presentation.

team offers an additional layer of protection against experimenter bias. However, obtaining blinded listeners' ratings of speech can be a time-consuming process, and many researchers would hesitate to incorporate outside listeners in a time-sensitive context such as this study. McAllister Byun et al. (2015) suggested that the process of obtaining blinded listener ratings of speech could be made rapid and predictable using online crowdsourcing platforms such as Amazon Mechanical Turk (AMT). Although online data collection tends to be "noisier" than data collection in the lab setting, aggregating responses over a larger number of individuals can improve the signal-to-noise ratio (Ipeirotis, Provost, Sheng, & Wang, 2014). Simulations reported in McAllister Byun et al. (2015) indicated that the "industry standard" level of agreement with an expert listener gold standard was matched when responses were aggregated across samples of at least nine AMT listeners. The present study followed that model by collecting binary ratings of each speech token from nine unique listeners recruited through AMT. This use of AMT to obtain speech ratings was approved by the Institutional Review Board at New York University, and all participants and parents of participants in treatment gave consent for sound files to be shared with external listeners in an anonymized fashion for rating purposes.

The minimum baseline duration was set to four sessions. After four sessions, the intervention team uploaded all participants' probe data for rating, and the analysis team downloaded and plotted the results. The analysis team could see a continuous trajectory of performance for each individual in the study, but they had no information about each individual's treatment status. Although both word- and syllable-level data were made available to the analysis team, the focus in making decisions about baseline stability was on vocalic /r/ at the syllable level, as noted above. Word-level data were consulted primarily as a source of potentially disambiguating information when syllable-level results were unclear. If any participant showed a rising trend or an outlier, the analysis team would direct the intervention team to collect two additional data points, then upload those data for rating and further review. If stability was regained, the analysis team would give the direction to initiate treatment for one randomly selected participant; otherwise, they would request that two additional data points be collected. If stability was still not achieved after a total of eight baseline sessions, treatment was initiated for one randomly selected participant in accordance with a time-based criterion. This provision was necessary in light of the abovementioned fact that most studies of biofeedback intervention for residual speech errors do include cases of nonresponders; it is important to avoid a situation where participants are held indefinitely at baseline while a nonresponsive participant receives treatment.

Table 1 reports operational definitions that were established through consultation between the intervention team and the analysis team in the example study. Although we consider it good practice to agree on estimates of these parameters prior to the initiation of data collection, we note that it can be challenging to specify exact values without a priori knowledge of how variable participants' scores will be at baseline or how rapidly they will respond once treatment is initiated. The analysis team may find it necessary to alter certain criteria in order to arrive at what they consider the most valid visual analysis of the plotted data.

An example of the determination to transition a participant from the baseline to the treatment phase is illustrated in Figure 3a with data from the first cohort of participants, who received ultrasound biofeedback treatment. Recall that the minimum baseline duration was four sessions; if all four participants showed a sufficiently stable baseline at that point, the analysis team would convey that treatment could be initiated for a randomly selected participant. However, one participant (pseudonym Katherine) exhibited a high outlier data point in the fourth baseline session. The analysis team thus requested two additional data points. Review of the additional two observations indicated that adequate stability of the data had been regained, so the analysis team communicated that treatment could be initiated for a randomly selected participant.

Once treatment had been initiated for one participant, the minimum stagger was set to four sessions, that is, 2 weeks of treatment; after that point, the intervention team processed and uploaded probe data from the previous four sessions from all participants to be rated by blinded listeners on AMT. The analysis team looked for one participant to show evidence of a treatment effect while the other participants maintained a stable baseline level of performance. If multiple participants were showing evidence of improvement, or if an outlier was observed, the same protocol used to extend the baseline phase (collect two more sessions, analyze again, repeat if necessary) was followed until a clear pattern of response emerged, or until the time-based criterion of eight sessions was reached.

An example of the determination to transition an additional participant from the baseline to the treatment phase is illustrated with data from Cohort 2 (visual-acoustic biofeedback) in Figure 3b. For an extended period after the transition from baseline to treatment, no participant showed a clear trajectory of change over time. The analysis team accordingly requested additional data points until a well-defined trend emerged. In the eighth session after the transition, one participant, Samantha, showed strong evidence of a treatment effect. The analysis team accordingly signaled that another participant could be randomly selected to receive treatment.

Lastly, a time-based criterion was the primary determinant of the end of the study: Participants completed 10 weeks (20 sessions) of intervention and then proceeded to complete three post-treatment maintenance probes. Note that in this fixed-duration design, if data points from one participant ceased to appear while others continued, the analysis team could deduce which participant was the first to receive treatment. This means that the intervention team needs to transmit dummy data for any participants who have completed treatment. In the present case, the intervention team recycled maintenance probes from the participants who had finished treatment (i.e., uploaded the data

Table 1. Operational definitions set for the sample study.

Term	Definition
Stable baseline	A series of at least four consecutive data points in which the most recent two data points do not demonstrate *evidence of improvement* or any *problematic outliers*.
Evidence of improvement	A participant will be judged to show significant evidence of improvement if: 1. The mean across sessions in the current phase, excluding the first two data points, is more than 10 percentage points higher than the mean in the preceding phase. 2. Treatment probes in the current phase show an upward trend over time (positive overall slope and final datapoint is at least 10 percentage points higher than the mean in the preceding phase).
Problematic outlier	An observation falling greater than two standard deviations above the mean across preceding data points.
Time-based criterion	If no participant shows a clear pattern of improvement after eight sessions, apply the time-based criterion and advance one randomly selected participant to treatment.

to be rerated by blinded listeners on AMT) until the time-based criterion was reached for all participants.

In the final step of the MVA procedure for multiple-baseline data, the analysis team reviewed all plotted data and made their best guess of the order in which participants received treatment. The final plotted data from the second cohort (visual-acoustic biofeedback) are represented in Figure 4. The study period is divided into six phases, with each boundary between phases representing a point at which the analysis team indicated that another participant could make the transition from baseline to treatment. The phases are not equal in duration across participants because of occasional absences that could not be made up before an upload for data rating took place. (As long as the analysis team keeps track of what session each participant was in when a phase transition occurred, these differences in phase duration have minimal impact on interpretation.)

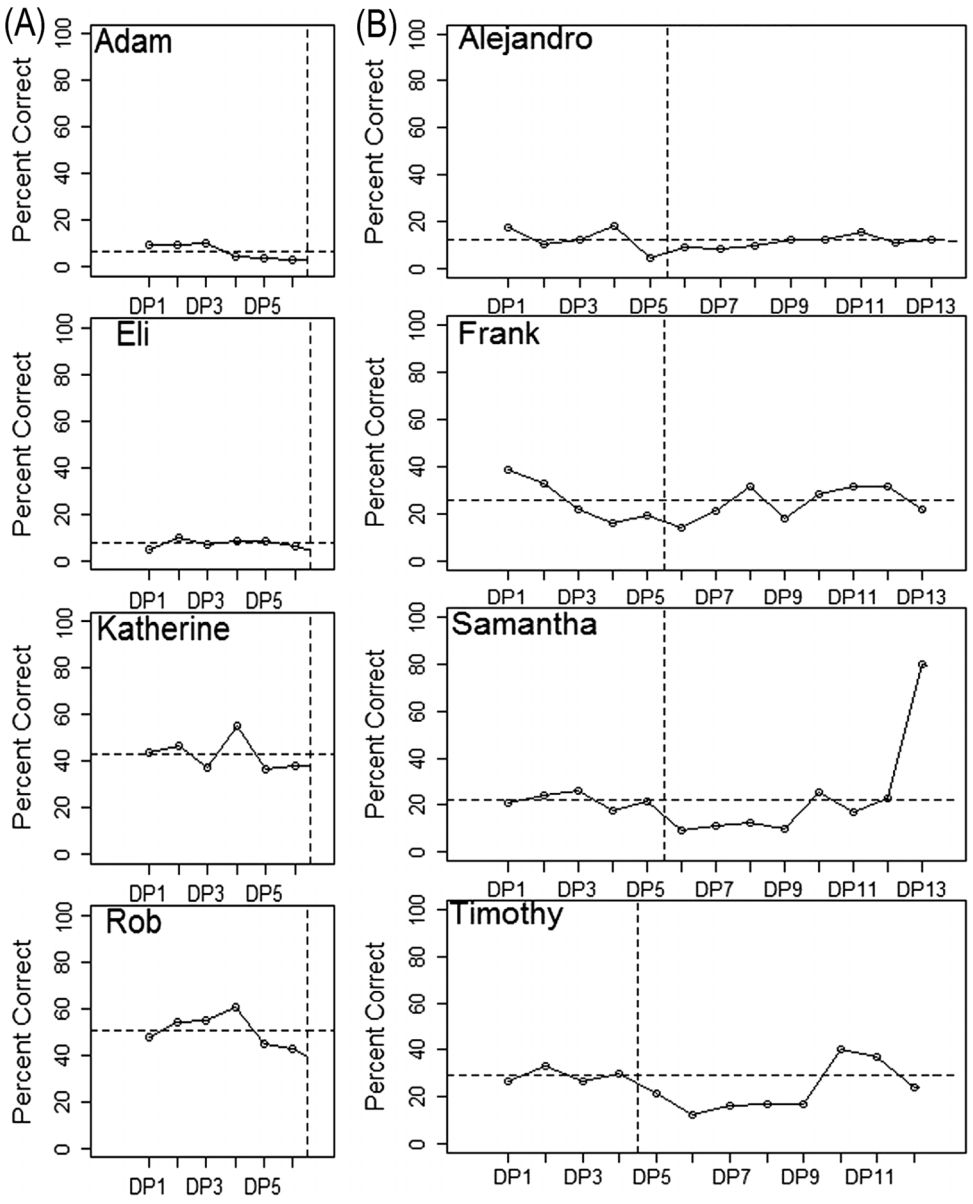
The first participant, Alejandro, showed a well-defined response pattern: His accuracy remained extremely stable in Phases 1–4 and then began to rise in Phase 5. From this, it was hypothesized that Alejandro was the fourth subject to receive treatment. The second participant, Frank, was unquestionably showing a treatment effect in Phase 4. Due to high variability in the early phases, there was some ambiguity about whether his response began in Phase 2 or 3. Comparison with data from word-level probes (not pictured) supported the hypothesis that his treatment effect began in Phase 3, and Frank was thus speculated to be the second subject to receive treatment. The third participant, Samantha, was judged to exhibit a treatment effect starting in Phase 2, as discussed above. Therefore, Samantha was hypothesized to be the first participant to receive treatment. The final participant, Tim, showed no response to treatment. However, based on their guesses regarding the other three participants, the analysis team was able to hypothesize that Tim was the third participant to receive treatment, and that the complete order was (1) Samantha, (2) Frank, (3) Tim, (4) Alejandro. The intervention team confirmed that this order was correct. With four participants, the total number of possible orderings was 4! (= 24), and the probability of correctly guessing the correct order was 1/24 = .04. Because this is an outcome that would be unlikely to occur by chance if the null hypothesis of no effect of treatment were true, the null hypothesis was provisionally rejected.

Discussion

This study reviewed a number of possible benefits that can be derived from the incorporation of MVA methods into single-case research in communication disorders. The MVA approach makes it possible to enjoy some of the major benefits of a response-guided approach—namely the ability to mitigate the impact of rising trends, problematic outliers, and other threats to internal validity—without introducing an excessive potential for experimenter bias to influence the outcomes of the study. In addition, MVA makes it possible to conduct a hypothesis test and compute a p value in conjunction with single-case data, contrasting with the purely qualitative assessment of outcomes that is used in many single-case studies. The case study reported here illustrates how MVA could be deployed in the context of a single-case study investigating treatment for residual errors affecting /r/, which is known to be a challenging context for obtaining interpretable single-case data. In the example provided, the analysis team was able to guess the correct order of treatment application, an outcome that was unlikely to occur by chance if the null hypothesis of no treatment effect were true ($p < .05$). This type of concrete result, paired with the higher internal validity of a blinded study, has the potential to enhance the credibility of single-case research. Although the present tutorial focused on the application of MVA in the context of a multiple-baseline across-subjects study, see Ferron and Levin (2014) for discussion of its use in the context of other designs.

Of course, the MVA method has disadvantages as well. The primary difficulty arises from the fact that this method is labor intensive, at least in the manner that it was implemented for this study. The authors found it challenging to operationalize all relevant parameters prior to the initiation of the study. For example, it was difficult to dictate what magnitude of progress should be considered meaningful when we did not yet have estimates of effect size for all technologies (e.g., EPG). In such cases, research teams may need to preserve some flexibility to adjust parameters during the study. Some of the complexity that characterized this study could be reduced by eliminating

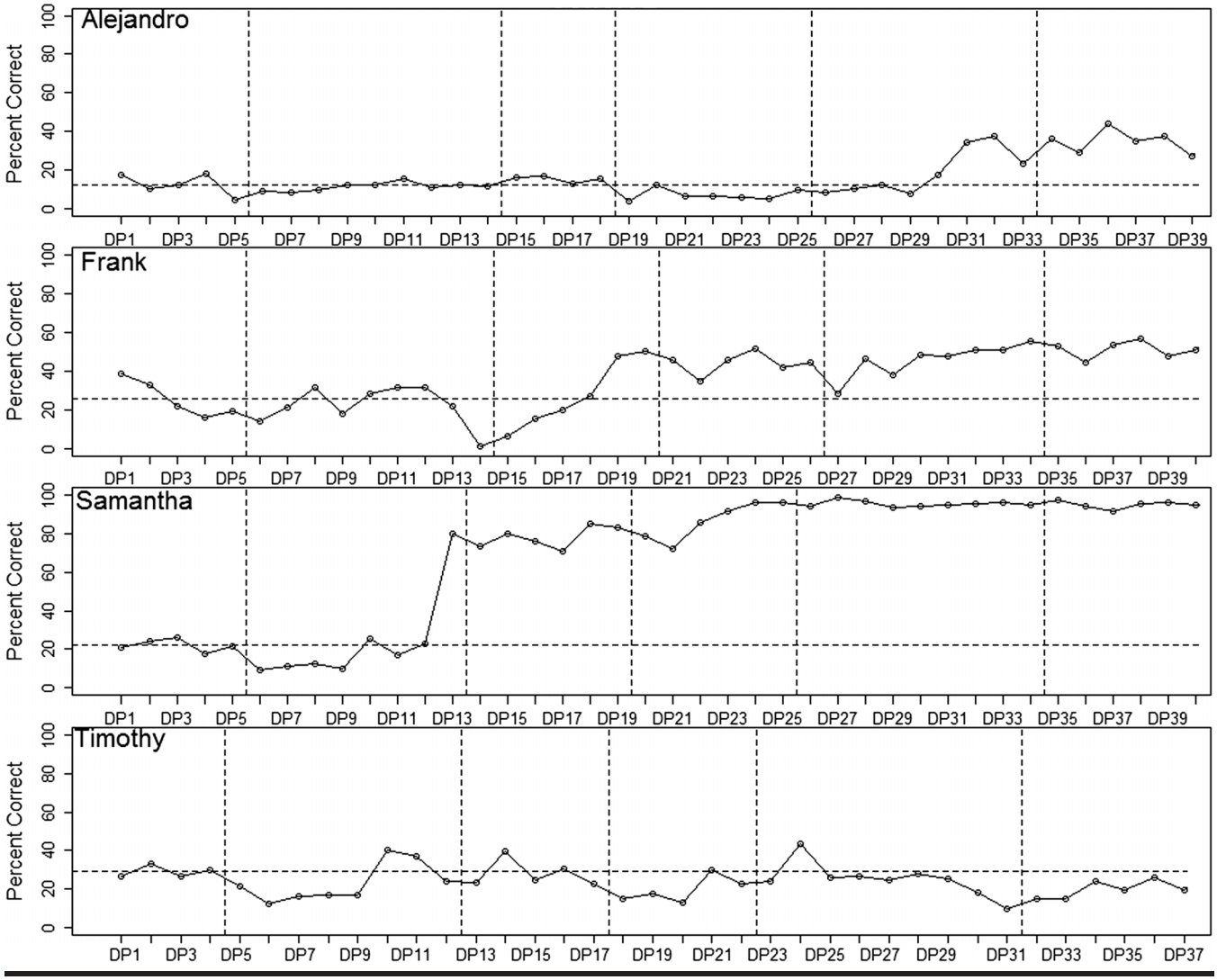
Figure 3. Schematic data illustrating decision points in a masked visual analysis design. *y* axis = percent of syllable-level tokens rated correct by blinded listeners. Vertical dotted lines represent boundaries between phases of the study. Horizontal dotted lines represent baseline mean for each participant. DP = data point. a: Example of establishing a stable baseline (data from Cohort 1). b: Example of identifying a response to treatment (data from Cohort 2).



our use of a third party (blinded raters on AMT) to provide ratings. However, if we eliminate this level of blinding, we reintroduce a real potential for bias to influence the outcome of the study: The intervention team may unconsciously give higher scores to participants receiving treatment than participants for whom treatment has not yet begun.

An additional set of limitations pertains to the scheduling needs of participating families. In our example study, the duration of the baseline phase ended up being as long as 28 sessions for the last child to transition from baseline to treatment. (Recall that all participants had the opportunity to receive the same duration of treatment regardless of the order in which they transitioned from baseline to treatment.) This is a much longer baseline phase than we would ever impose outside of the context of the MVA model. Parents need to be amply prepared for this contingency. In our experience, after a conversation with the research team about the importance of controlled studies (and the value of free therapy), most families are willing to participate even with the knowledge that their child could be the last to enter treatment. The situation becomes more complicated if families need to travel some distance to reach the lab for baseline data collection. In

Figure 4. Schematic data illustrating the process of guessing overall treatment order (data from Cohort 2). The *y*-axis represents percent of syllable-level tokens rated correct by blinded listeners. Vertical dotted lines represent boundaries between phases of the study. Horizontal dotted line represents baseline mean for each participant. DP = data point.



some cases, we found it necessary to send students to participants' homes to collect baseline recordings or to train parents to collect recordings in a quiet room using a portable digital recorder borrowed from the lab. Although such recordings are of lower quality than would be obtained in a sound booth, it is up to the researcher to strike a balance between optimizing the quality of the data and proposing a protocol that is feasible for the family as well as the research team.

Concerns can also be raised regarding the power of MVA to detect effects. In our study, which had four subjects, there were 24 possible orders and thus the order had to be correctly guessed on the first try to obtain a statistically significant result (i.e., $p < .05$). Had there been multiple nonresponders, it is unlikely that the order would have been correctly guessed; however, not detecting an effect in such circumstances is consistent with single-case guidelines that suggest minimum evidence of treatment effectiveness requires three demonstrations of the effect at three points in time (Kratochwill et al., 2013). In multiple-baseline studies with only three subjects, there are only six possible orders, and thus to have any power to detect effects, the randomization scheme has to be altered. Instead of randomly ordering the three subjects, the intervention team could randomly select without replacement from the set {Subject 1, Subject 2, Subject 3, and no one} at each point of phase transition, yielding 24 possible assignments (Ferron & Levin, 2014). The statistical power of MVA in multiple-baseline studies has been examined using simulation algorithms that approximate the decision making of masked visual analysts. These simulation studies showed that Type I errors in MVA were sufficiently controlled in both fixed phase length and response-guided designs, and that the power was higher for the response-guided design than it was

for the fixed phase length design (Ferron, Joo, & Levin, in press).

A final concern, which is general to hypothesis testing, is that statistically significant results may be incorrectly interpreted as clinically significant results. To help limit these sorts of misinterpretations it is recommended that the significance test from the MVA not be interpreted in isolation, but rather as one component in the analysis. For example, the hypothesis test in the MVA could be used to formally control for Type I errors, and if the null hypothesis is rejected, a traditional visual analysis could be used to assess the clinical significance of the effect (Ferron et al., in press).

Conclusions

In sum, MVA is an approach to single-case data collection that, although it has drawbacks, in many ways represents a "best-of-both-worlds" combination. It does require discipline and detailed preparation by both the intervention and analysis teams, as well as cooperative families. A major benefit is that data analysis and plotting are completed in the course of conducting the study—once treatment is complete, the study is in effect ready to be written up. We found it worth the effort to incorporate MVA into our study comparing treatments for residual /r/ errors, which pose a unique challenge for single-case design research. Finally, we perceive the adoption of MVA as compatible with a broader shift toward better-designed intervention studies in the field of communication disorders. The defining features of MVA—thorough operationalization of study parameters, rigorous blinding, and randomization—are also critical prerequisites for internal validity in experimental research. Although these elements require additional effort and attention from the researcher, they also promise to yield higher quality evidence on the important subject of treatment efficacy.

Acknowledgments

This research was supported by NIH R03DC 012883 (McAllister Byun) and also benefited from a travel fellowship to attend the IES Single-Case Design and Analysis Institute 2014 (McAllister Byun). The authors also express their gratitude to the following individuals: for study implementation and data management, Roberta Lazarus, Lauren Dioguardi, and Melissa Lopez; for programming support, José Ortiz and Daniel Szeredi; and for comments on the article, graduate students in NYU's reading group on single-case design for communication disorders. Many thanks as well to all participants (including online participants) and their families for their cooperation throughout the study.

References

Beeson, P. M., & Robey, R. R. (2006). Evaluating single-subject treatment research: Lessons learned from the aphasia literature. *Neuropsychology Review, 16*(4), 161–169.

Brossart, D. F., Parker, R. I., Olson, E. A., & Mahadevan, L. (2006). The relationship between visual analysis and five statistical analyses in a simple AB single-case research design. *Behavior Modification, 30*(5), 531–563.

Busk, P. L., & Serlin, R. C. (1992). Meta-analysis for single-case research. In T. R. Kratochwill & J. R. Levin (Eds.), *Single case research design and analysis: New directions for psychology and education* (pp. 187–212). Hillsdale, NJ: Erlbaum.

Byiers, B. J., Reichle, J., & Symons, F. J. (2012). Single-subject experimental design for evidence-based practice. *American Journal of Speech-Language Pathology, 21*(4), 397–414.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Culton, G. L. (1986). Speech disorders among college freshmen: A 13-year survey. *Journal of Speech and Hearing Disorders, 51*(1), 3–7.

Ferron, J., & Jones, P. K. (2006). Tests for the visual analysis of response-guided multiple-baseline data. *The Journal of Experimental Education, 75*(1), 66–81.

Ferron, J. M., Joo, S.-H., & Levin, J. R. (in press). A Monte-Carlo evaluation of masked-visual analysis in response-guided versus fixed-criteria multiple-baseline designs. *Journal of Applied Behavior Analysis.*

Ferron, J. M., & Levin, J. R. (2014). Single-case permutation and randomization statistical tests: Present status, promising new developments. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case intervention research: Methodological and statistical advances* (pp. 153–183). Washington, DC: American Psychological Association.

Fletcher, S. G., Dagenais, P. A., & Critz-Crosby, P. (1991). Teaching consonants to profoundly hearing-impaired speakers using palatometry. *Journal of Speech, Language, and Hearing Research, 34*(4), 929–943.

Gierut, J. A., & Morrisette, M. L. (2011). Effect size in clinical phonology. *Clinical Linguistics and Phonetics, 25*(11-12), 975–980.

Gierut, J. A., Morrisette, M. L., & Dickinson, S. L. (2015). Effect size for single-subject design in phonological treatment. *Journal of Speech, Language, and Hearing Research, 58*(5), 1464–1481.

Hitchcock, E. R., & McAllister Byun, T. (2015). Enhancing generalisation in biofeedback intervention using the challenge point framework: A case study. *Clinical Linguistics and Phonetics, 29*(1), 59–75.

Hitchcock, E. R., McAllister Byun, T., Swartz, M. T., & Lazarus, R. (in press). Effectiveness of electropalatography for treating misarticulation of /r/. *American Journal of Speech-Language Pathology.*

Horner, R. H., & Odom, S. L. (2014). Constructing single-case research designs: Logic and options. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case intervention research: Methodological and statistical advances* (pp. 27–52). Washington, DC: American Psychological Association.

Howard, D., Best, W., & Nickels, L. (2015). Optimising the design of intervention studies: Critiques and ways forward. *Aphasiology, 29*(5), 526–562.

Ipeirotis, P. G., Provost, F., Sheng, V. S., & Wang, J. (2014). Repeated labeling using multiple noisy labelers. *Data Mining and Knowledge Discovery, 28*(2), 402–441.

Kazdin, A. E. (2010). *Single-case research designs: Methods for clinical and applied settings* (2nd ed.). New York, NY: Oxford University Press.

Kearns, K. P. (1986). Flexibility of single-subject experimental designs. Part II: Design selection and arrangement of experimental phases. *Journal of Speech and Hearing Disorders, 51*(3), 204–214.

Klein, H. B., McAllister Byun, T., Davidson, L., & Grigos, M. I. (2013). A multidimensional investigation of children's /r/

productions: Perceptual, ultrasound, and acoustic measures. *American Journal of Speech-Language Pathology, 22*(3), 540–553.

Kratochwill, T. R., Hitchcock, J. H., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2013). Single-case intervention research design standards. *Remedial and Special Education, 34*(1) 26–38.

Kratochwill, T. R., & Levin, J. R. (2010). Enhancing the scientific credibility of single-case intervention research: Randomization to the rescue. *Psychological Methods, 15*(2), 124.

Kratochwill, T. R., & Levin, J. R. (2014). *Single-case intervention research: Methodological and statistical advances.* Washington, DC: American Psychological Association

Maas, E., Robin, D. A., Austermann Hula, S. N., Freedman, S. E., Wulf, G., Ballard, K. J., & Schmidt, R. A. (2008). Principles of motor learning in treatment of motor speech disorders. *American Journal of Speech-Language Pathology, 17,* 277–298.

Matyas, T. A., & Greenwood, K. M. (1990). Visual analysis of single-case time series: Effects of variability, serial dependence, and magnitude of intervention effects. *Journal of Applied Behavior Analysis, 23*(3), 341–351.

McAllister Byun, T., Halpin, P. F., & Szeredi, D. (2015). Online crowdsourcing for efficient rating of speech: A validation study. *Journal of Communication Disorders, 53,* 70–83.

McAllister Byun, T., & Hitchcock, E. R. (2012). Investigating the use of traditional and spectral biofeedback approaches to intervention for /r/ misarticulation. *American Journal of Speech-Language Pathology, 21*(3), 207–221.

McAllister Byun, T., Hitchcock, E. R., & Ortiz, J.. (2014). Challenge-R: Computerized challenge point treatment for /r/ misarticulation. Talk presented at the Annual Convention of the American Speech-Language-Hearing Association, Orlando, FL.

McAllister Byun, T., Hitchcock, E. R., & Swartz, M. T. (2014). Retroflex versus bunched in treatment for rhotic misarticulation: Evidence from ultrasound biofeedback intervention. *Journal of Speech, Language, and Hearing Research, 57*(6), 2116–2130.

McAllister Byun, T., Swartz, M. T., Halpin, P. F., Szeredi, D., & Maas, E. (2016). Direction of attentional focus in biofeedback treatment for /r/ misarticulation. *International Journal of Language and Communication Disorders, 51*(4), 384–401.

McReynolds, L. V., & Kearns, K. (1983). *Single-subject experimental designs in communicative disorders.* Austin, TX: Pro-Ed.

McReynolds, L. V., & Thompson, C. K. (1986). Flexibility of single-subject experimental designs. Part I: Review of the basics of single-subject designs. *Journal of Speech and Hearing Disorders, 51*(3), 194–203.

OCEBM Levels of Evidence Working Group. (2011). *The Oxford 2011 Levels of Evidence*. Retrieved from http://www.cebm.net/index.aspx?o=5653

Preston, J. L., Brick, N., & Landi, N. (2013). Ultrasound biofeedback treatment for persisting childhood apraxia of speech. *American Journal of Speech-Language Pathology, 22*(4), 627–643.

Preston, J. L., McCabe, P., Rivera-Campos, A., Whittle, J. L., Landry, E., & Maas, E. (2014). Ultrasound visual feedback treatment and practice variability for residual speech sound errors. *Journal of Speech, Language, and Hearing Research, 57*(6), 2102–2115.

Robey, R. R. (2004). A five-phase model for clinical-outcome research. *Journal of Communication Disorders, 37*(5), 401–411.

Robey, R. R., Schultz, M. C., Crawford, A. B., & Sinner, C. A. (1999). Review: Single-subject clinical-outcome research: Designs, data, effect sizes, and analyses. *Aphasiology, 13*(6), 445–473.

Ruscello, D. M. (1995). Visual feedback in treatment of residual phonological disorders. *Journal of Communication Disorders, 28,* 279–302.

Rvachew, S. (1988). Application of single subject randomization designs to communicative disorders research. *Human Communication Canada, 12*(4), 713.

Rvachew, S., & Brosseau-Lapré, F. (2012). *Developmental phonological disorders: Foundations of clinical practice*. San Diego, CA: Plural Publishing.

Scruggs, T. E., Mastropieri, M. A., & Casto, G. (1987). The quantitative synthesis of single-subject research methodology and validation. *Remedial and Special Education, 8*(2), 24–33.

Shriberg, L. D. (2010). Childhood speech sound disorders: From postbehaviorism to the postgenomic era. In R. Paul & P. Flipsen (Eds.), *Speech sound disorders in children* (pp. 1–34). San Diego, CA: Plural Publishing.

Shriberg, L. D., Tomblin, J. B., & McSweeny, J. L. (1999). Prevalence of speech delay in 6-year-old children and comorbidity with language impairment. *Journal of Speech, Language, and Hearing Research, 42*(6), 1461–1481.

Thompson, C. K. (2006). Single subject controlled experiments in aphasia: The science and the state of the science. *Journal of Communication Disorders, 39*(4), 266–291.

Wolery, M., Busick, M., Reichow, B., & Barton, E. E. (2010). Comparison of overlap methods for quantitatively synthesizing single-subject data. *The Journal of Special Education, 44*(1), 18–28.